TD n° 8 : Exercices de synthèse

Exercice 1. On considère une population U constituée de 4 individus : $U = \{u_1, u_2, u_3, u_4\}$. On prélève 3 individus dans U suivant un plan de sondage de type **PE avec remise** (et non pas PEAR comme traité en cours) formant ainsi un échantillon. **PE avec remise** signifie que tous les individus ont la même probabilité d'être sélectionné (d'où le **PE**) et qu'un individu peut être sélectionné plusieurs fois (d'où le **avec remise**).

1. Combien d'échantillons peut-on former (l'ordre des individus étant pris en compte) ?
2. Quelle est la probabilité d'obtenir un échantillon ne contenant pas u_4 ?

Exercice 2. Un jour donné, un fast-food comptent 200 clients. Sur un échantillon de 18 clients prélevé suivant un plan de sondage aléatoire de type PESR, les dépenses, en euros, sont :

21.35	15.60	12.10	26.20	35.80	18.75	21.40	35.15	22.65

19.25	17.30	15.70	22.05	21.80	17.70	20.55	17.15	23.65

1. Calculer le taux de sondage.
2. Donner une estimation ponctuelle de la moyenne des dépenses des 200 clients.
3. Donner une estimation ponctuelle de l'écart-type de l'estimateur de la moyenne des dépenses des 200 clients.
4. Déterminer un intervalle de confiance au niveau 95% pour la moyenne des dépenses des 200 clients.
5. Déterminer la taille d'échantillon à choisir pour avoir une incertitude relative sur la moyenne des dépenses des 200 clients inférieure ou égale à 1% au niveau 95%.

Exercice 3. Dans une population de 3 individus $U = \{u_1, u_2, u_3\}$, on sélectionne au hasard et sans remise 2 individus pour former un échantillon. La *var* W égale à l'échantillon obtenu vérifie :

$$\mathbb{P}(u_1 \in W) = \frac{2}{10}, \qquad \mathbb{P}(u_2 \in W) = \frac{9}{10}, \qquad \mathbb{P}(W = \{u_1, u_2\}) = \frac{1}{10}.$$

Dans l'immédiat, il n'y a pas plus d'information sur u_3.

1. Déterminer, sans calcul, $\mathbb{P}(W \in \{u_1, u_2, u_3\})$.
2. Déterminer, sans calcul, $\mathbb{P}(\{u_1 \in W\} \cup \{u_2 \in W\})$. Retrouver ce résultat à l'aide d'une formule de probabilités et de calculs.
3. Pour tout $i \in \{1, 2, 3\}$, exprimer simplement $\mathbb{P}(u_i \in W)$ en fonction de $\mathbb{P}(W = \{u_i, u_j\})$ et $\mathbb{P}(W = \{u_i, u_k\})$, avec $j, k \in \{1, 2, 3\}/\{i\}$ et $j \neq k$. Sans utiliser les valeurs des probabilités mises en jeu, en déduire la valeur de

$$\mathbb{P}(u_1 \in W) + \mathbb{P}(u_2 \in W) + \mathbb{P}(u_3 \in W).$$

4. Calculer $\mathbb{P}(u_3 \in W)$.
5. On admet que $\mathbb{P}_{\{u_3 \in W\}}(u_1 \in W) = 1/9$. Calculer $\mathbb{P}(W = \{u_1, u_3\})$, puis $\mathbb{P}(W = \{u_2, u_3\})$.
6. Calculer $\mathbb{P}_{\{u_2 \in W\}}(u_3 \in W)$.

Exercice 4.

1. Créer un tableau `dataframe` de 300 lignes que l'on appellera `Enquete` avec trois colonnes:
   - une colonne `Campus` avec 3 modalités : `C1`, `C2` et `C3`, répétées chacune 185, 70 et 45 fois respectivement,
   - une colonne `HF` avec 2 modalités: `H` et `F` répétées chacune 150 fois et réparties au hasard et uniformément,
   - une colonne `Lecture` avec 2 modalités : `0` et `1` réparties au hasard avec une probabilité de 35% d'avoir `1`.
2. Créer une nouvelle data frame appelée `Enquete2` qui contient toutes les lignes de `Enquete` vérifiant `C1` ou `C2` dans la colonne `Campus`.
3. Construire un vecteur de 300 charactères appelé `X` dont les entrées correspondent aux couples de modalités `C1-H`, `C1-F`, `C2-H` et `C2-F` de chaques lignes de `Enquete2`.
4. Expliquer les enjeux des commandes suivantes:

```
n = tapply(Enquete2$Lecture, list(X), sum)
p = tapply(Enquete2$Lecture,list(X), sum) / sum(n)
```

Exercice 5. On considère le jeu de données islands. Les données sont disponibles ici :

```
library(datasets)
data(islands)
islands
```

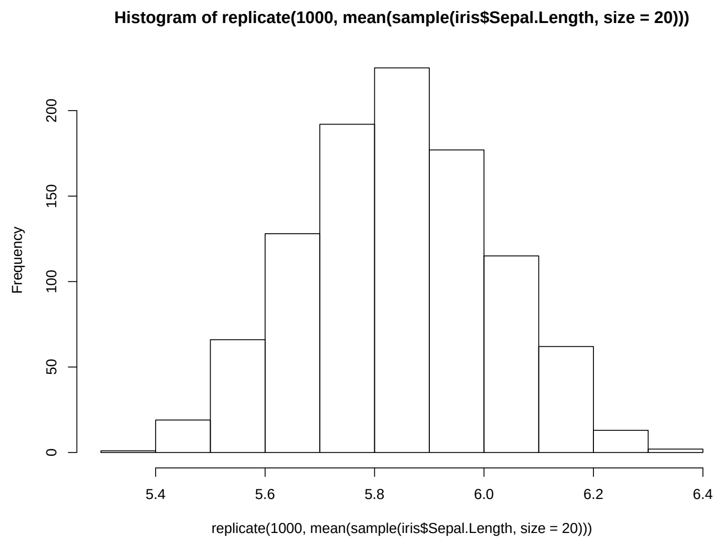
Écrire les commandes R permettant de sélectionner un échantillon de 8 îles suivant un plan de sondage aléatoire de type PESR de deux manière différentes:

- avec la commande sample,
- avec la commande srswor,
- avec la commande slice_sample.

Exercice 6. On fait les commandes R suivantes :

```
hist(replicate(1000, mean(sample(iris$Sepal.Length, size = 20))))
```

Cela renvoie :



Expliquer l'enjeu de ces commandes ainsi que la sortie graphique.

Exercice 7.

On fait les commandes R suivantes :

```
w1 = rnorm(1000)
w2 = rnorm(1000, 2, 2)
w3 = rnorm(1000, 1, 3)
w = cbind(w1, w2, w3)
n = nrow(w)
nb = floor(nrow(w) * 0.75)
s = sample(1:n, nb)
ap = w[s, ]
te = w[-s, ]
```

1. Donner la valeur de `n` et la valeur de `nb`.
2. Donner l'entier logiquement proche de `mean(w[,2])`.
3. Donner l'entier logiquement proche de `sd(w[,3])`.
4. Qu'est-ce que `ap` par rapport à `w` ?
5. Qu'est-ce que `te` par rapport à `w` et à `ap` ?
6. Est-ce que `rbind(ap, te)` correspond exactement à `w` ?
7. Que renvoie les commandes : `sum(rbind(ap, te)) == sum(w)` ?

Exercice 8. Dans un premier temps, charger la librairie `tidyverse`. Il s'agit de répondre aux question suivante en utilisant un maximum de fonctions de la librairie `dplyr`.

1. Créer le tableau `tibble` suivant:

```
## # A tibble: 9 x 6
##   name   school teacher sex    math_score reading_score
##   <chr>  <chr>  <chr>   <chr>       <dbl>         <dbl>
## 1 mike   south  johnson male            4             1
## 2 carol  south  johnson female          3             5
## 3 greg   south  johnson male            2             2
## 4 marcia south  johnson female          4             4
## 5 peter  north  smith   male            3             5
## 6 jan    north  smith   female          4             4
## 7 bobby  north  smith   male            5             1
## 8 cindy  south  perry   female          4             5
## 9 alice  south  perry   female          5             4
```

2. Trier les données en fonction des valeurs de `math_score`, de la plus haute à la plus basse. Qui a eu le meilleur score en maths ?

3. Créer un jeu de données contenant uniquement les colonnes: `name`, `math_score` et `reading_score`.

4. Sélectionner les individus vérifiant `male` et `south`.

5. Les notes reportées dans `math_score` et `reading_score` sont sur 5. Remplacer ces colonnes par des colonnes contenant des notes équivalentes mais sur 20 (ainsi, `mike` a 16 (sur 20) au `math_score` et 4 (sur 20) au `reading_score`, etc.).

6. Déterminer la valeur minimale de `math_score` pour les deux modalités de `school` (on fera une analyse groupée).

7. Créer un jeu de données constitué de 4 individus sélectionnés avec un plan de sondage de type PESR, et des valeurs des caractères associées.

Exercice 9. Que peut-on dire du graphique suivant ?